**Slavin Chapter 10: Measurement**

**What is Reliability and Validity?**

Reliability –degree to which measures produce consistent results—goal is to show same score or result with different populations—Range from 0 (No Reliability) to 1 (Perfect Reliability)—Reliability places an upper limit on the degree to which it can correlate with anything else—using measure that are low in reliability increases the chance of failing to detect true relationships or making false negatives

* Scales- test and questionnaire type reliabilities are usually computed on scales
* Internal Consistency (split-half, KR20, alpha coefficient)—degree to which scores on items in a scale correlate with one another
* Interrater Reliability—ratio of agreement s between two or more observers or raters to the total number of observations
* Test-retest—the correlation between two or more observers or raters to the total number of observations

Validity—the degree to which it actually measures the concept that it is supposed to measure—important for research design—measures may be reliable but not have validity—there is no numerical value for validity(it is either valid or not valid)

* Face validity-degree to which a measure appears to assess what it is supposed to asses
* Content validity—degree to which test items correspond to the content of a course, etc.
* Predictive validity—degree to which scores on a scale or test predict later behaviors or scores
* Concurrent validity—degree to which a scale or test correlate with another conceptually related scale, test, or other variable measured at the same time
* Construct validity-the degree to which a scale or test has a pattern of correlation with other variable that would be predicted by sound theory

**What are Types of Measures?**

1. Achievement and Aptitude Tests
   1. Aptitude tests- measure potential for learning given knowledge or skills
   2. Achievement tests- measure actual acquisition of knowledge and skills
      1. Standardized tests-carefully constructed tests and used on nationally representative samples to enable the test developers to establish norms (Norm-referenced test) to indicate how test taker compares to all students
      2. Criterion-referenced tests- built around a well-defined set of instructional objectives—focus on student scores rather than a comparison to a national norm—designed to be high in content validity—may use alternative to multiple choice formats
         1. Authentic tests—form of criterion-referenced test that is performance based—carry out a task
         2. Researcher-made, Content-specific tests—more closely covers content
         3. Problems—Researcher made, content specific—little use as a pretest due to 0 scores
   3. Questionnaire Scales—must be shown to measure reliably and validly –using existing scales is better than attempting to create a new one—existing scales are often normed, usually have reliability and validity established, and help tie studies to previous literature
   4. Behavior Observation—observation in an environment for a) High Incident behaviors- require observer to use a great deal of judgment [Ex. rate someone on warmth, activeness, or enthusiasm] or b) Low Incident behaviors-require less judgment by the observer [ex. time on task, number of questions answered per student per class]
      1. Constructing A Behavioral Observation System
         1. How will you define each observation category?
         2. How will you schedule your observation sessions?
         3. Within an observation session, whom will you observe?
         4. What schedule will you use?
      2. Reliability and Bias—unreliable behavioral observation can create bias—very specific definitions and directions must be written and followed to ensure reliability—avoid bias by not giving observers the hypothesis or a stake in the study (co-authorship)—Reliability can be increased with observer practice
         1. Overall Reliability—indicates the percentage of all observations on which both observers saw the same thing
         2. Occurrence Reliability—a measure of reliability that compares the number of observation intervals in which each two observers agreed divided by the number of intervals

\*\*when we wish to use a test as a covariate, reliability is critical and standardized or criterion referenced tests are the best –they correlate better with other measures

**How do you determine Sample Size? –**depends on how variable the data are and on how large we expect the effect to be—each additional subject adds to the ability to find a true effect if it is there or to be confident if no effect exists—Additional subjects also reduce **Statistical Power** (the ability of a research design to detect true differences and avoid false negative errors) –as the sample size gets larger it takes an even larger increase to get further increases in statistical power